A User Interface to Communicate Interpretable AI Decisions to Radiologists

Yanchen Jessie Ou[†,a], Alina Jade Barnett[†,a], Anika Mitra[†,a], Fides Regina Schwartz[b], Chaofan Chen[c], Lars Grimm[b], Joseph Y. Lo[b,d,e], and Cynthia Rudin[a,e]

[a]Department of Computer Science, Duke University, Durham, USA
[b]Department of Radiology, Duke University, Durham, USA
[c]School of Computing and Information Science, University of Maine, Orono, USA
[d]Department of Biomedical Engineering, Duke University, Durham, USA
[e]Department of Electrical and Computer Engineering, Duke University, Durham, USA
[†]Denotes equal contribution

ABSTRACT

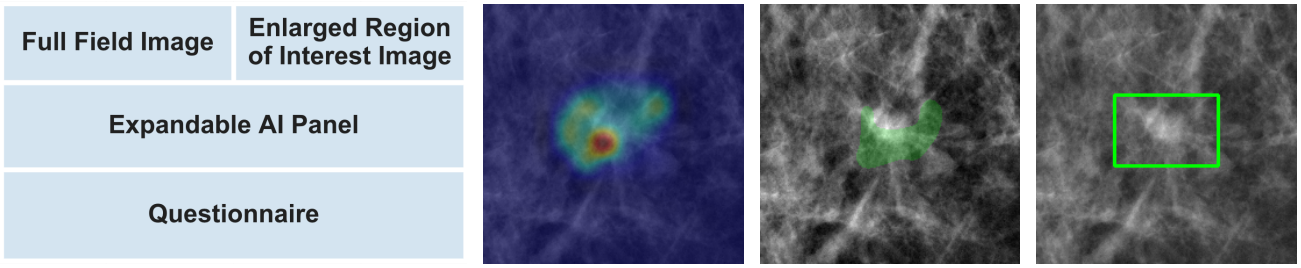
Tools for computer-aided diagnosis based on deep learning have become increasingly important in the medical field. Such tools can be useful, but require effective communication of their decision-making process in order to safely and meaningfully guide clinical decisions. Inherently interpretable models provide an explanation for each decision that matches their internal decision-making process. We present a user interface that incorporates the Interpretable AI Algorithm for Breast Lesions (IAIA-BL) model, which interpretably predicts both mass margin and malignancy for breast lesions. The user interface displays the most relevant aspects of the model's explanation including the predicted margin value, the AI confidence in the prediction, and the two most highly activated prototypes for each case. In addition, this user interface includes full-field and cropped images of the region of interest, as well as a questionnaire suitable for a reader study. Our preliminary results indicate that the model increases the readers' confidence and accuracy in their decisions on margin and malignancy.

Keywords: Interpretability, Deep Learning, Communication, User Interface, Breast, Mammography, Risk Assessment, Cancer

1. INTRODUCTION & PURPOSE

The prevalent use of deep learning has become transformative in many fields, offering impressive speed and accuracy in analyzing data and predicting diagnoses. Black box machine learning models are common, but should not be used in high-stakes decisions given their susceptibility to bias and incomprehensibility.[1] To this end, many researchers have proposed methods to attempt to build trust in these models. Such interpretation methods, like saliency maps, CAM, and GradCAM, provide some insight into how black-box models work, but have significant performance issues[2–4] and thus, are not necessarily reliable.[5,6] On the other hand, newer works show it is both possible[7,8] and more helpful[1] to implement neural networks that are inherently interpretable – these models display the rationale for arriving at a specific decision. Many researchers[9–11] have identified interpretability as a major barrier to the clinical deployment of deep learning models. Specifically in a medical context, common radiologic tasks such as lesion detection, and diagnosis and prognosis are high-stakes practices with significant consequences, requiring transparent and useful computer aid. Additionally, current techniques in interpretable machine learning often prioritize technical feasibility over user-design.[12] Thus, the advent of deep transparent models that show the reasoning behind their predictions offer us a unique opportunity to better understand what information is clinically relevant, and how to build trust in high-stakes algorithms.

Technology from interpretable machine learning[13] and explainable artificial intelligence (XAI)[14,15] can be used in computer-vision applications, commonly in radiology to guide clinical decisions. Thus it is crucial that these models explain their decision-making process with detail and precision. Previously, Barnett et al.[16] created Interpretable AI Algorithm for Breast Lesions (IAIA-BL), which uses an interpretable mass margin classification model to predict the margin and malignancy of a breast lesion. Because this interpretable model provides explanations at different levels of detail, we built a user interface to bridge the gap between machine learning



(a) (b) (c) (d)

Figure 1: (a) Layout of the user interface for the four-block retrospective reader study. See Appendix A for a screen capture of the entire interface. (b-d) Three methods to show the AI activation over the input image, in order of decreasing information. (b) Heat-map. (c) Semi-transparent covered overlay. (d) Bounding Box.

experts and radiologists by presenting a simplified explanation, while allowing the reader to request a more detailed explanation when desired.

2. NEW WORK

The user interface development and description is newly introduced in this paper. Previous submissions of IAIA-BL have demonstrated its effectiveness as an interpretable model using mass margin and mass shape,[17] while this paper includes such work in a graphical user interface suitable for radiologists in a clinical setting.

3. METHODS

Our user interface displays a full-field image digital mammogram scan with the region of interest highlighted in a bounding box alongside a cropped ROI image. Underneath these images is an AI panel that shows the prediction margin, the model's confidence, and buttons to open explanation image panels. The user interface includes a questionnaire panel that uses sliding percentage scales and free-comment text boxes to gauge the readers' confidence and accuracy in their decisions. A view of the entire user interface is shown in Appendix A.

3.1 AI Panel Development

We developed various formats for displaying information in the explanations panel of the user interface. While original explanations of IAIA-BL employed heat maps, as shown in Figure 1b, these presented too much information to be quickly understood by a radiologist. We observed that bounding box images as presented in Figure 1d did not reveal enough information to be considered useful in a clinical setting, as they would include too much of the lesion. Hence, we updated the user interface to display images with semi-transparent covered overlay indicating the top 5% of pixels from the heat map. This is shown in Figure 1c and allows radiologists to easily identify the similar regions. As there are both merits and drawbacks of including the colored overlay on the images for the explanations panel, we implemented two buttons that open and close separate panels that can be configured to the readers' liking, as presented in Figure 2. Finally, as activation scores produced by the model were raw floating point numbers, we recast these numerical values into qualitative similarity bins (high, medium, and low), to be more easily understood by a user.

3.2 Data & Model

The model used for UI development is IAIA-BL, described by Barnett et al.[8] This model is based on the interpretable deep computer-vision architecture ProtoPNet,[7] with changes made to apply to digital mammography. This model uses a prototype method to make predictions entirely based on the similarity between the test image and learned prototypes. For a set of learned prototypes, the explanations are of the form: "this part of the input image looks like that part from a learned prototype." This method yields an explanation with perfect fidelity, as the reasoning presented in the explanation exactly matches the reasoning the model uses to make its decisions. Our data set includes 1136 images of breast masses from Duke University Health Systems, selected from patients who received breast mass biopsies between 2008 and 2018. Using the associated clinical reports, we labelled for

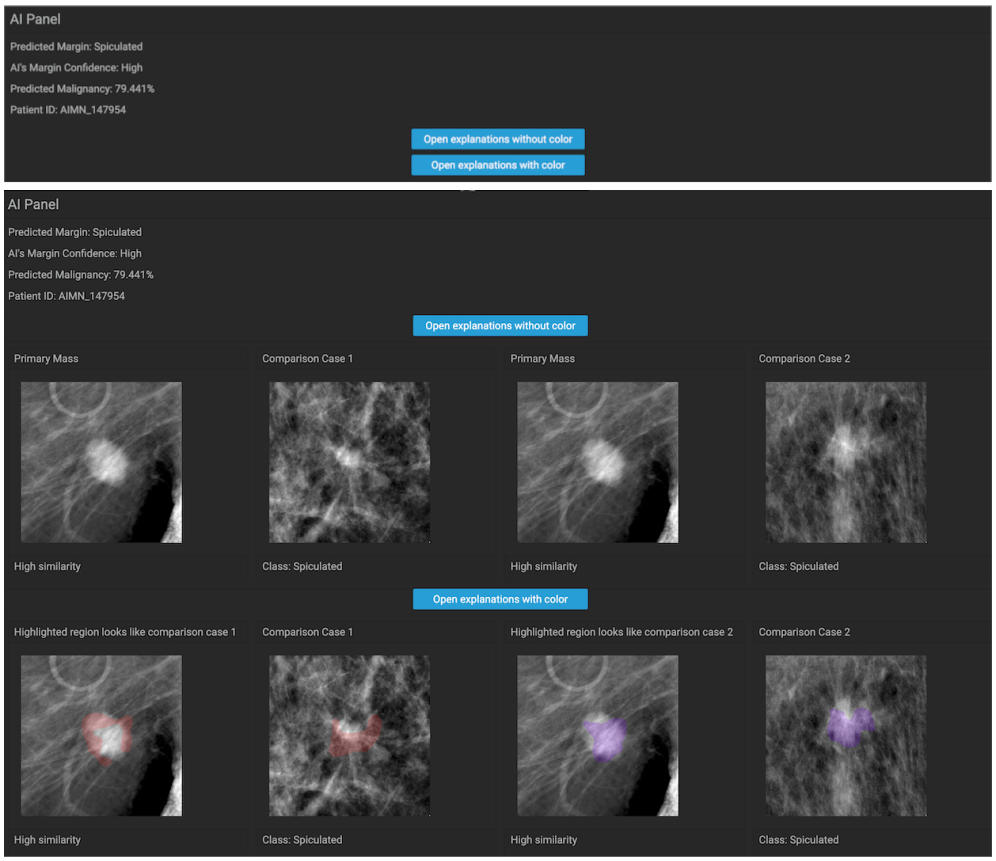


Figure 2: Top: AI panel with both buttons closed, showing only minimal information (AI prediction and confidence). Bottom: AI panel fully expanded, showing both the explanations with color and the explanations without color. Both panels include headings, as well as labels to identify similarity of comparison case to primary mass and class of comparison cases. A user may choose to expand either panel individually or expand and view both panels.

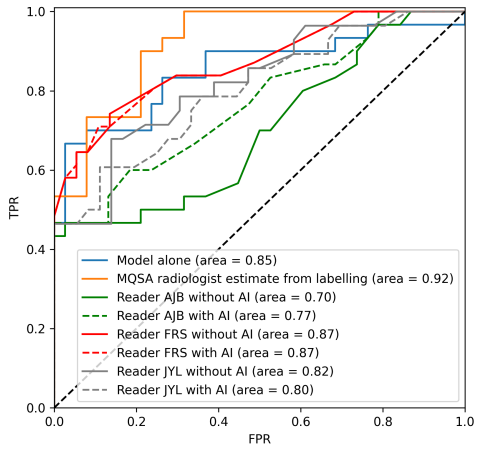


Figure 3: The ROC curves of malignancy prediction for the model alone, the original physician-labellers (fellowship-trained MQSA breast imaging radiologists), and the trial readers both with and without AI assistance.

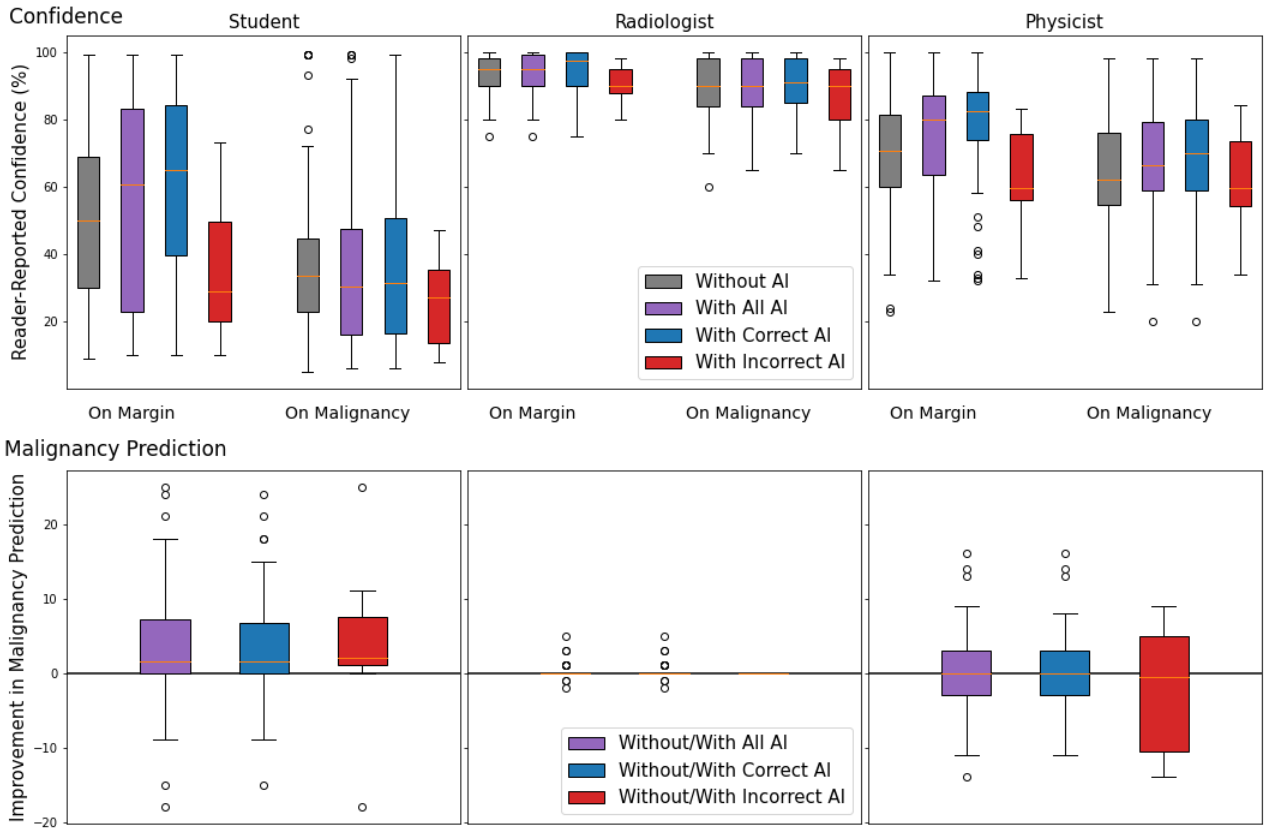


Figure 4: Results from the trial run of the user study. Top: Reader-reported prediction confidence. Bottom: Change in reader's estimate of the malignancy score. Note: Reader FS (radiologist) rarely changed their malignancy score prediction, so the boxes overlap with the 0 line.

malignancy using histopathology results as ground truth. Each ROI is individually labeled for mass margin and mass shape by one breast radiologist (LG) with 7 years of experience.

4. RESULTS

The fully-developed user interface will be used in a larger four-block retrospective reader study in the coming months to analyze the usefulness of our AI tool. In two blocks, the users will be shown the user interface with the AI panel outlined above and in the other two blocks, without.

We performed a trial run with three readers: a radiologist (FRS) with 9 post-graduation years of general clinical experience, a medical physics professor (JYL) with 20 years of experience, and a graduate student in Computer Science with 3 years experience working with mammographic images (AJB). The study includes the 75 patients that will be assessed during the reader study. The results indicate that AI assistance improved reader confidence for all readers when the AI was correct. With AI assistance, reader malignancy predictions improved for the student regardless of whether the AI was correct, but worsened for the physicist when the AI was incorrect. Use of the AI assistance made no change for the radiologist malignancy predictions. However, this is only a very small trial run, so no general conclusions can be made.

Furthermore, we obtained qualitative data from our readers in response to the study. The readers reported unanticipated examples of self-activation on circumscribed prototypes – instead of activating at the margins alone, the AI activated the entire lesion. At times, the AI explanations seemed contradictory – it output low similarity scores when comparison cases were highly similar to the test case visually, thereby confusing the radiologist readers. In addition, a number of cases had their top two most similar comparison cases drawn from

the same mammographic image or the same patient. This qualitative feedback proved that we have successfully interfaced the AI and the clinicians: now, clinicians have adequate understanding of the algorithm to be able to comment directly on the underlying models. The feedback will be incorporated in future development to improve weakness in the AI model itself, by having better coverage of margins, and drawing prototypes from different patients. This type of feedback was possible because the AI model is transparent: it would be extremely difficult to obtain this type of feedback for improving black box models, which tend to have the same problems as interpretable deep learning models but are much more difficult to troubleshoot. A large-scale reader study with more radiologists and trainees could follow such improvements.

5. CONCLUSIONS

We aim to address the need for a tool that bridges the interpretable model IAIA-BL and viewing its results in a clinical setting. Leveraging our collaboration of specialists in interpretable machine learning and clinical experts, we addressed this problem by designing and building a comprehensible and accessible user interface. After updating the interface to include the desired features, it can be used in further reader studies to obtain results on the efficacy of our machine learning models in a pseudo-clinical setting.

ACKNOWLEDGMENTS

We would like to acknowledge breast radiologists Michael Taylor-Cho, Lars Grimm, Connie Kim, and Sora Yoon, who annotated the dataset, as well as Joe Shamblin who helped us with the back-end implementation of our UI.

We acknowledge the Duke MEDx High-Risk High-Reward grant and the NSF HRD-2222336 grant for funding.

REFERENCES

[1] Rudin, C., "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence* **1**, 206–215 (2019).

[2] Nguyen, G., Kim, D., and Nguyen, A., "The effectiveness of feature attribution methods and its correlation with automatic evaluation scores," *arXiv preprint arXiv:2105.14944* (2021).

[3] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B., "Sanity checks for saliency maps," *Advances in Neural Information Processing Systems* **31** (2018).

[4] Draelos, R. L. and Carin, L., "Hirescam: Faithful location representation in visual attention for explainable 3d medical image classification," *arXiv preprint arXiv:2011.08891* (2020).

[5] Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M. D., and Kalpathy-Cramer, J., "Assessing the (un)trustworthiness of saliency maps for localizing abnormalities in medical imaging," *medRxiv* (2020).

[6] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L., "Explaining Explanations: An Overview of Interpretability of Machine Learning," (2018).

[7] Chen, C., Li, O., Tao, C., Barnett, A., Su, J., and Rudin, C., "This looks like that: Deep learning for interpretable image recognition," in [*Proceedings of Neural Information Processing Systems (NeurIPS)*], (2019).

[8] Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., and Rudin, C., "A case-based interpretable deep learning model for classification of mass lesions in digital mammography," *Nature Machine Intelligence* **3**(12), 1061–1070 (2021).

[9] Langlotz, C. P., Allen, B., Erickson, B. J., Kalpathy-Cramer, J., Bigelow, K., Cook, T. S., Flanders, A. E., Lungren, M. P., Mendelson, D. S., Rudie, J. D., Wang, G., and Kandarpa, K., "A roadmap for foundational research on artificial intelligence in medical imaging: From the 2018 nih/rsna/acr/the academy workshop," *Radiology* **291**(3), 781–791 (2019).

[10] Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F.-M., Tengg-Kobligk, H. v., Summers, R. M., and Wiest, R., "On the interpretability of artificial intelligence in radiology: Challenges and opportunities," *Radiology: Artificial Intelligence* **2**, e190043 (2020).

[11] Prevedello, L. M., Halabi, S. S., Shih, G., Wu, C. C., Kohli, M. D., Chokshi, F. H., Erickson, B. J., Kalpathy-Cramer, J., Andriole, K. P., and Flanders, A. E., "Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions," *Radiology: Artificial Intelligence* **1**(1), e180031 (2019).

[12] Chen, H., Gomez, C., Huang, C.-M., and Unberath, M., "Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review," *npj Digital Medicine* **5** (2022).

[13] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C., "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistics Surveys* **16**(none), 1 – 85 (2022).

[14] Yang, G., Ye, Q., and Xia, J., "Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Information Fusion* **77**, 29–52 (2022).

[15] Singh, A., Sengupta, S., and Lakshminarayanan, V., "Explainable deep learning models in medical image analysis," *Journal of Imaging* **6** (2020).

[16] Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., and Rudin, C., "Interpretable mammographic image classification using cased-based reasoning and deep learning," *Deep Learning, Case-Based Reasoning, and AutoML: Present and Future Synergies, Workshop of International Joint Conferences on Artificial Intelligence Organization* (2021).

[17] Barnett, A. J., Sharma, V., Gajjar, N., Fang, J., Schwartz, F., Chen, C., Lo, J. Y., and Rudin, C., "Interpretable deep learning models for better clinician-ai communication in clinical mammography," in [*Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment*], **12035**, SPIE (2022).

Appendix A

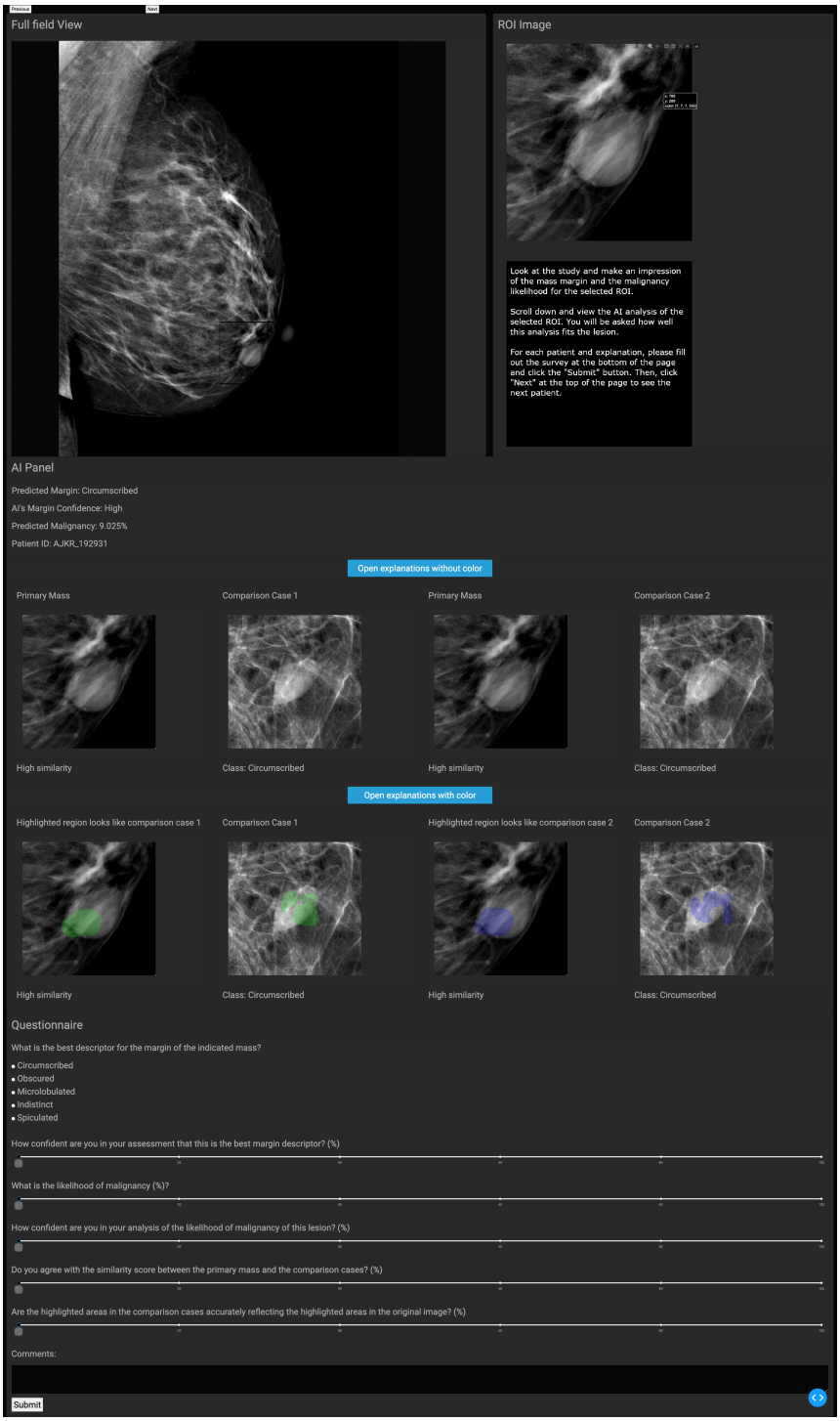


Figure 5: Long screenshot of the entire user interface for one patient. The top row shows the full-field image, region of interest cropped image, and an instructions image. The AI panel displays predicted margin and AI confidence, and two buttons that open explanations image panels. Finally the questionnaire assesses readers' confidence or accuracy in their decisions.